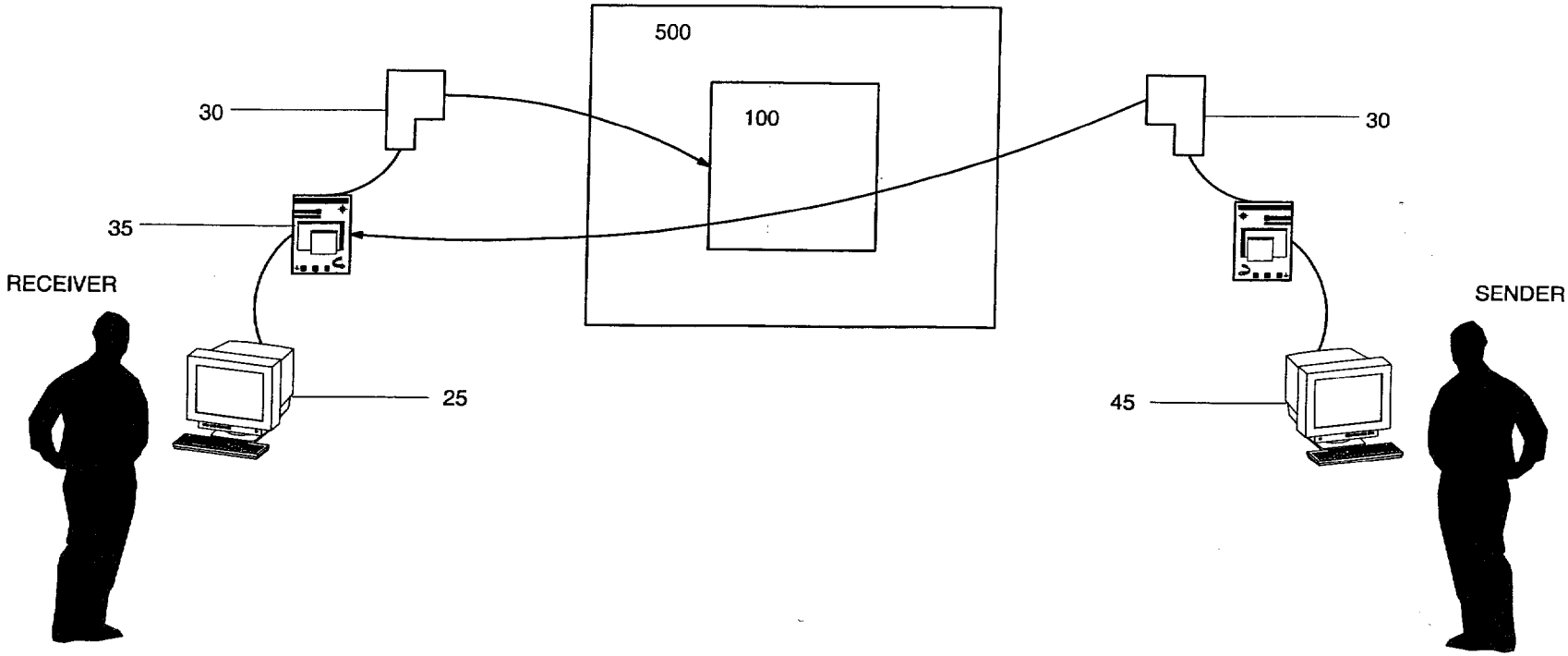

FIG. 1

First Name, Last Name Format	Extraction Algorithm States
Michael Smith	2,6
M. Smith	2,6
M. J. Smith	2,4,6
Michael J. Smith	2,4,6
M. Joseph Smith	2,4,6
Michael Joseph Smith	2,4,6
Mr. Smith	1,3
Mr. Michael Smith	1,2,6
Mr. M. Smith	1,2,6
Mr. M. J. Smith	1,2,4,6
Mr. Michael J. Smith	1,2,4,6
Mr. M. Joseph Smith	1,2,4,6
Mr. Michael Joseph Smith	1,2,4,6
Last Name, First Name Format	
Smith, Michael	3,5
Smith, M.	3,5
Smith, M. J.	3,5,7
Smith, Michael J.	3,5,7
Smith, M. Joseph	3,5,7
Smith, Michael Joseph	3,5,7
Smith, Mr.	3,3
Smith, Mr. Michael	3,3,5
Smith, Mr. M.	3,3,5
Smith, Mr. M. J.	3,3,5,7
Smith, Mr. Michael J.	3,3,5,7
Smith, Mr. M. Joseph	3,3,5,7
Smith, Mr. Michael Joseph	3,3,5,7
Post/Ancestor Combinations	
III	8
PhD	7
III PhD	8,7

Figure 2

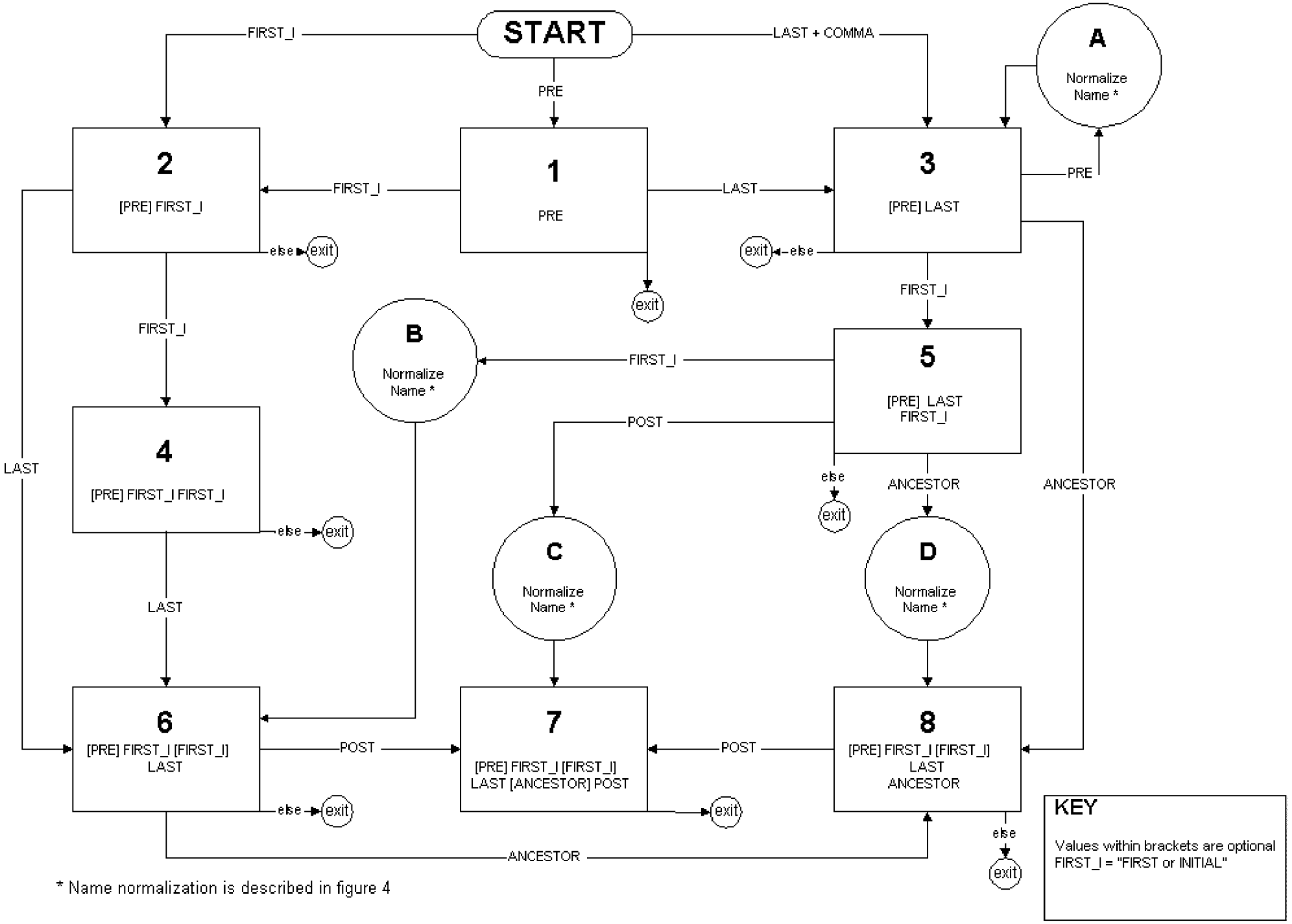
START
FIRST_I
LAST + COMMA
PRE
A
Normalize Name *
2
[PRE] FIRST_I
1
PRE
3
[PRE] LAST
FIRST_I
LAST
PRE
else
exit
exit
else
exit
FIRST_I
B
Normalize Name *
FIRST_I
5
[PRE] LAST FIRST_I
4
[PRE] FIRST_I FIRST_I
POST
LAST
else
exit
ANCESTOR
ANCESTOR
else
exit
C
Normalize Name *
D
Normalize Name *
LAST
LAST
6
[PRE] FIRST_I [FIRST_I] LAST
POST
7
[PRE] FIRST_I [FIRST_I] LAST [ANCESTOR] POST
POST
8
[PRE] FIRST_I [FIRST_I] LAST ANCESTOR
else
exit
exit
ANCESTOR
else
exit
KEY
Values within brackets are optional
FIRST_I = "FIRST or INITIAL"
* Name normalization is described in figure 4

Figure 3

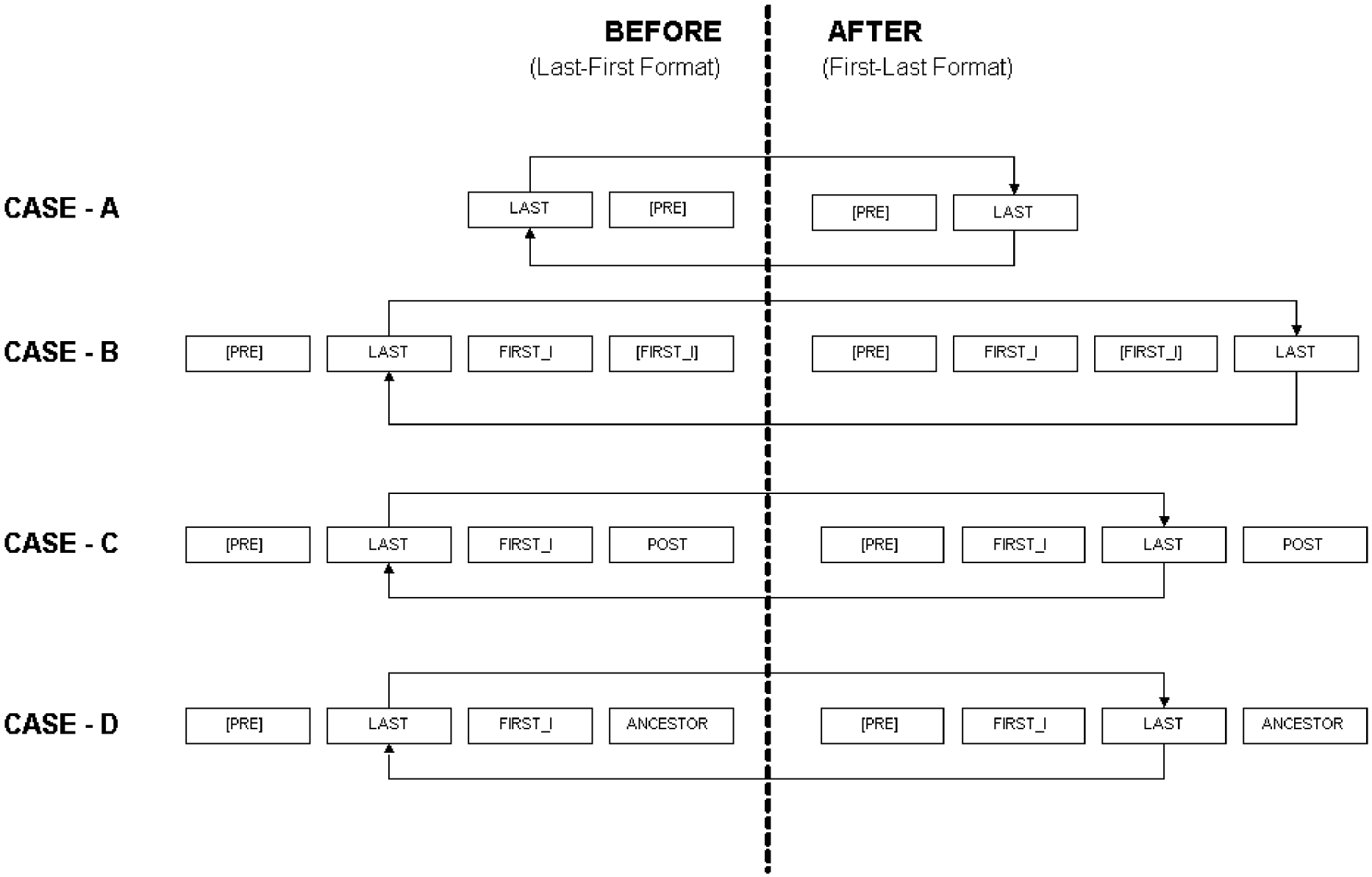
BEFORE
(Last-First Format)
AFTER
(First-Last Format)
CASE - A
LAST
[PRE]
[PRE]
LAST
CASE - B
[PRE]
LAST
FIRST_I
[FIRST_I]
[PRE]
FIRST_I
[FIRST_I]
LAST
CASE - C
[PRE]
LAST
FIRST_I
POST
[PRE]
FIRST_I
LAST
POST
CASE - D
[PRE]
LAST
FIRST_I
ANCESTOR
[PRE]
FIRST_I
LAST
ANCESTOR

Figure 4

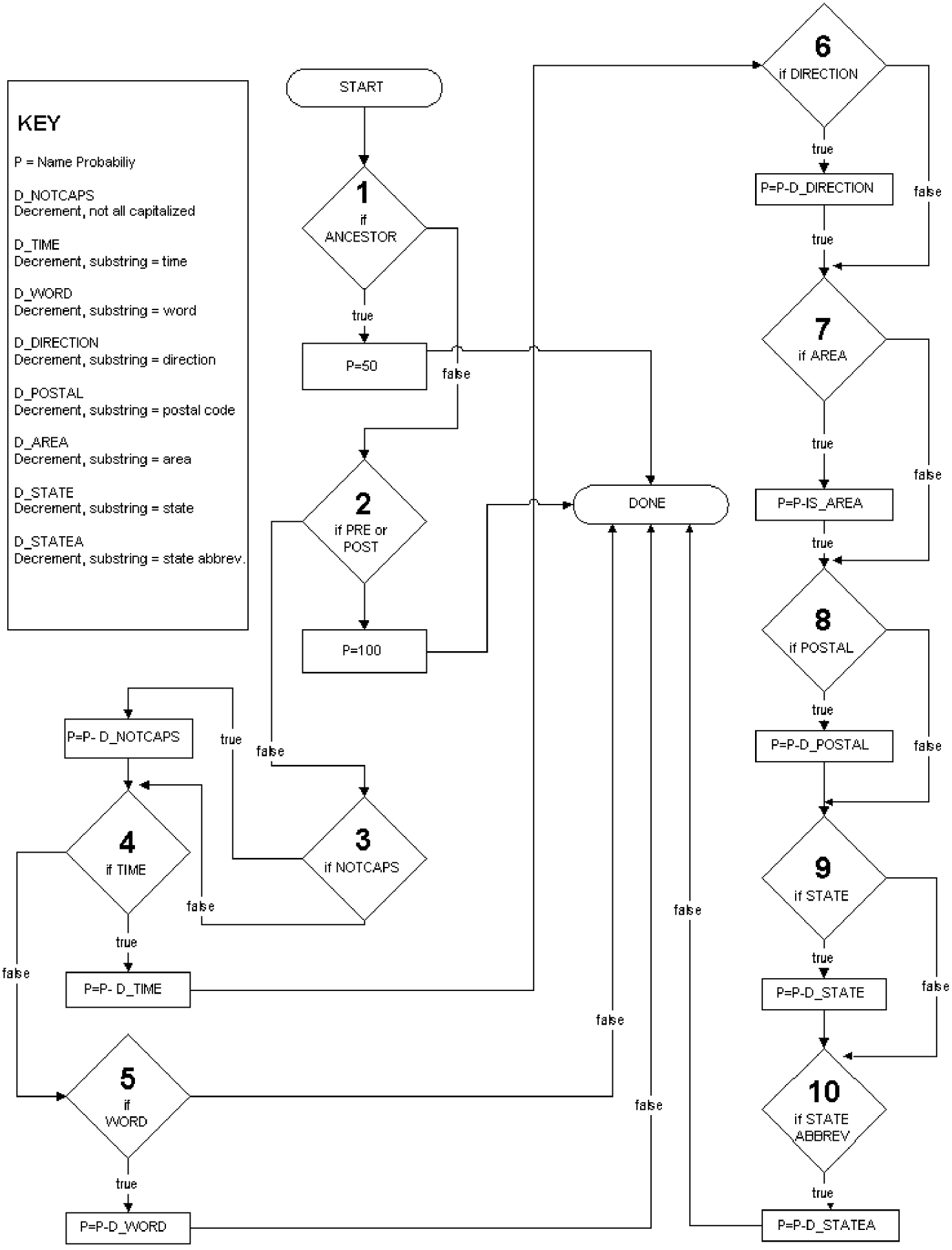
KEY
P = Name Probabiliy
D_NOTCAPS
Decrement, not all capitalized
D_TIME
Decrement, substring = time
D_WORD
Decrement, substring = word
D_DIRECTION
Decrement, substring = direction
D_POSTAL
Decrement, substring = postal code
D_AREA
Decrement, substring = area
D_STATE
Decrement, substring = state
D_STATEA
Decrement, substring = state abbrev.
START
1 if ANCESTOR
true
P=50
false
2 if PRE or POST
P=100
DONE
P=P- D_NOTCAPS
true
false
4 if TIME
3 if NOTCAPS
false
true
P=P- D_TIME
false
5 if WORD
true
P=P-D_WORD
false
false
6 if DIRECTION
true
P=P-D_DIRECTION
false
true
7 if AREA
true
P=P-IS_AREA
false
true
8 if POSTAL
true
P=P-D_POSTAL
false
9 if STATE
true
P=P-D_STATE
false
10 if STATE ABBREV
true
P=P-D_STATEA

Figure 5

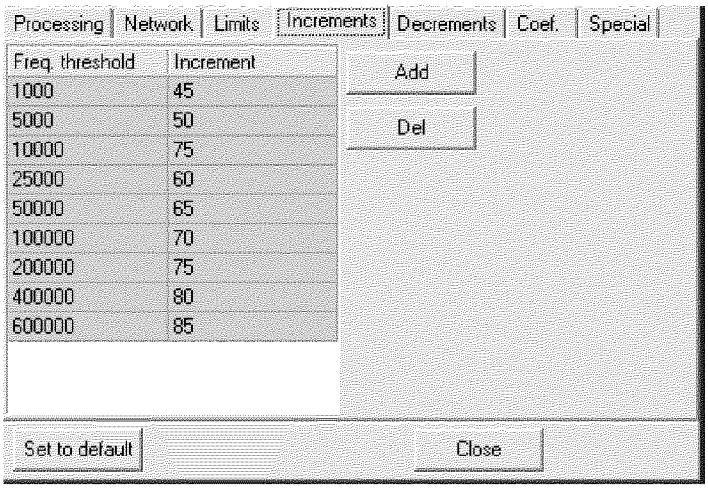
Processing
Network
Limits
Increments
Decrements
Coef.
Special
Freq. threshold
Increment
1000
45
5000
50
10000
75
25000
60
50000
65
100000
70
200000
75
400000
80
600000
85
Add
Del
Set to default
Close

Figure 6

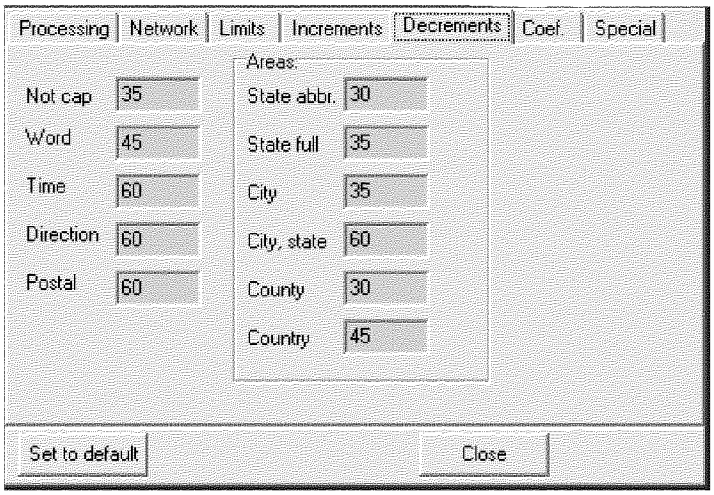
Processing
Network
Limits
Increments
Decrements
Coef.
Special
Areas:
Not cap
35
Word
45
Time
60
Direction
60
Postal
60
State abbr.
30
State full
35
City
35
City, state
60
County
30
Country
45
Set to default
Close

Figure 7

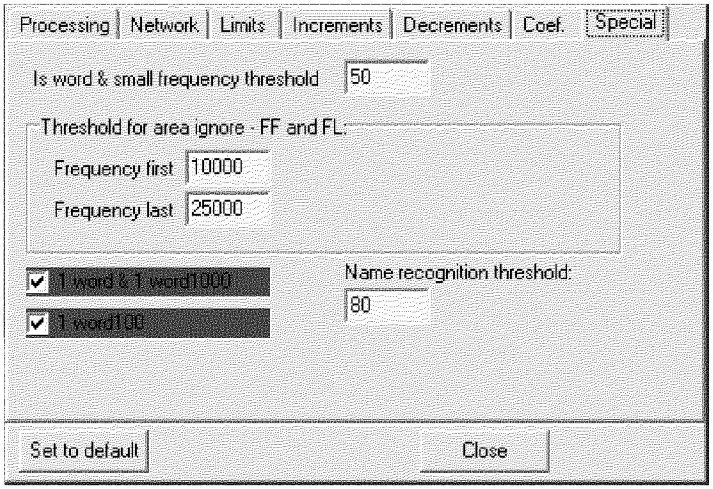
Processing
Network
Limits
Increments
Decrements
Coef.
Special
Is word & small frequency threshold
50
Threshold for area ignore - FF and FL:
Frequency first
10000
Frequency last
25000
1 word & 1 word1000
1 word100
Name recognition threshold:
80
Set to default
Close

Figure 8

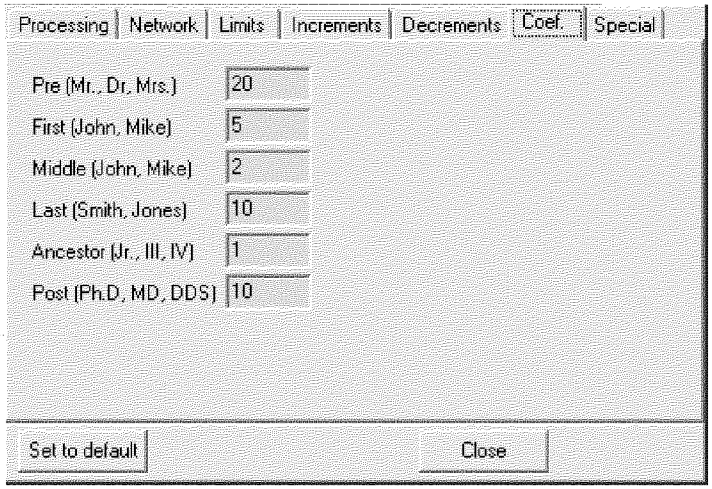
Processing
Network
Limits
Increments
Decrements
Coef.
Special
Pre (Mr., Dr, Mrs.)
20
First (John, Mike)
5
Middle (John, Mike)
2
Last (Smith, Jones)
10
Ancestor (Jr., III, IV)
1
Post (Ph.D, MD, DDS)
10
Set to default
Close

Figure 9

$$\frac{\sum_{i=1}^{6} X[i]\,K[i]\,P[i]}{\sum_{i=1}^{6} X[i]\,K[i]}$$

Figure 10

1. Values for X

X[1] = PRE	X[4] = LAST
X[2] = FIRST1	X[5] = ANCESTOR
X[3] = FIRST2	X[6] = POST

2. Values for K

K[1] = PRE	K[4] = LAST
K[2] = FIRST1	K[5] = ANCESTOR
K[3] = FIRST2	K[6] = POST

3. Values for P

P[1] = PRE	P[4] = LAST
P[2] = FIRST1	P[5] = ANCESTOR
P[3] = FIRST2	P[6] = POST

Figure 11

i	Name part	Substring	X	K	P
1	PRE	Mr	1	20	100
2	FIRST1	Donato	1	5	145
3	FIRST2	S	1	2	100
4	LAST	Diorio	1	10	145
5	ANCESTOR		0	1	0
6	POST		0	10	0

Figure 12

Initial equation with values filled in from example:

$$\frac{X[1]*K[1]*P[1] + X[2]*K[2]*P[2] + X[3]*K[3]*P[3] + X[4]*K[4]*P[4] + X[5]*K[5]*P[5] + X[6]*K[6]*P[6]}{X[1]*K[1] + X[2]*K[2] + X[3]*K[3] + X[4]*K[4] + X[5]*K[5] + X[6]*K[6]}$$

Completing the equation

1.

$$\frac{1*20*100 + 1*5*145 + 1*2*100 + 1*10*145 + 0*1*0 + 0*10*0}{1*20 + 1*5 + 1*2 + 1*10 + 0*1 + 0*10}$$

2.

$$\frac{2000 + 725 + 200 + 1450 + 0 + 0}{20 + 5 + 2 + 10 + 0 + 0}$$

3.

$$\frac{4375}{37}$$

4. Final name score for "Mr Donato S. Diorio" = 118.24

Figure 13

Source	Name	Context	Location
whois	Mark Tortorici	Domain Name: PROQWEST.COM Administrative Contact: **Tortorici, Mark** (MT3493) markt@PROQWEST.COM ProQwest 592 Weddell	
homepage	James Duran	Duran HCP Team **James Duran** President 408-540-0071 408-893-4905 jamesd@du	duranteam.html
homepage	Susan Fox	licon Valley recruiting today... More... **Susan Fox** Vice President of Operations 408-540-0072 408-	duranteam.html
homepage	Joel Abraham	ng and Life Science start ups... More... **Joel Abraham** Director of Midwest Operations 414-908-7799 41	duranteam.html
homepage	Jeff Polo	ograms to best meet their needs. More... **Jeff Polo** Client Services Manager 408-540-0073 408-893-4	duranteam.html
homepage	Bianca Medina	olunteer helping area schools... More... **Bianca Medina** Accounting & HR 408-540-0074 biancam@duran	duranteam.html
homepage	Peter Weddle	and employment site rankings are The Green Book by **Peter Weddle** , and CareerXroads, by Gerry Crispin and Mark Mehl	maynews.html
homepage	Gerry Crispin	e Green Book by Peter Weddle, and CareerXroads, by **Gerry Crispin** and Mark Mehler. Both do comprehensive jobs of ra	maynews.html
homepage	Mark Mehler	ter Weddle, and CareerXroads, by Gerry Crispin and **Mark Mehler** . Both do comprehensive jobs of ranking and classi	maynews.html

Figure 14